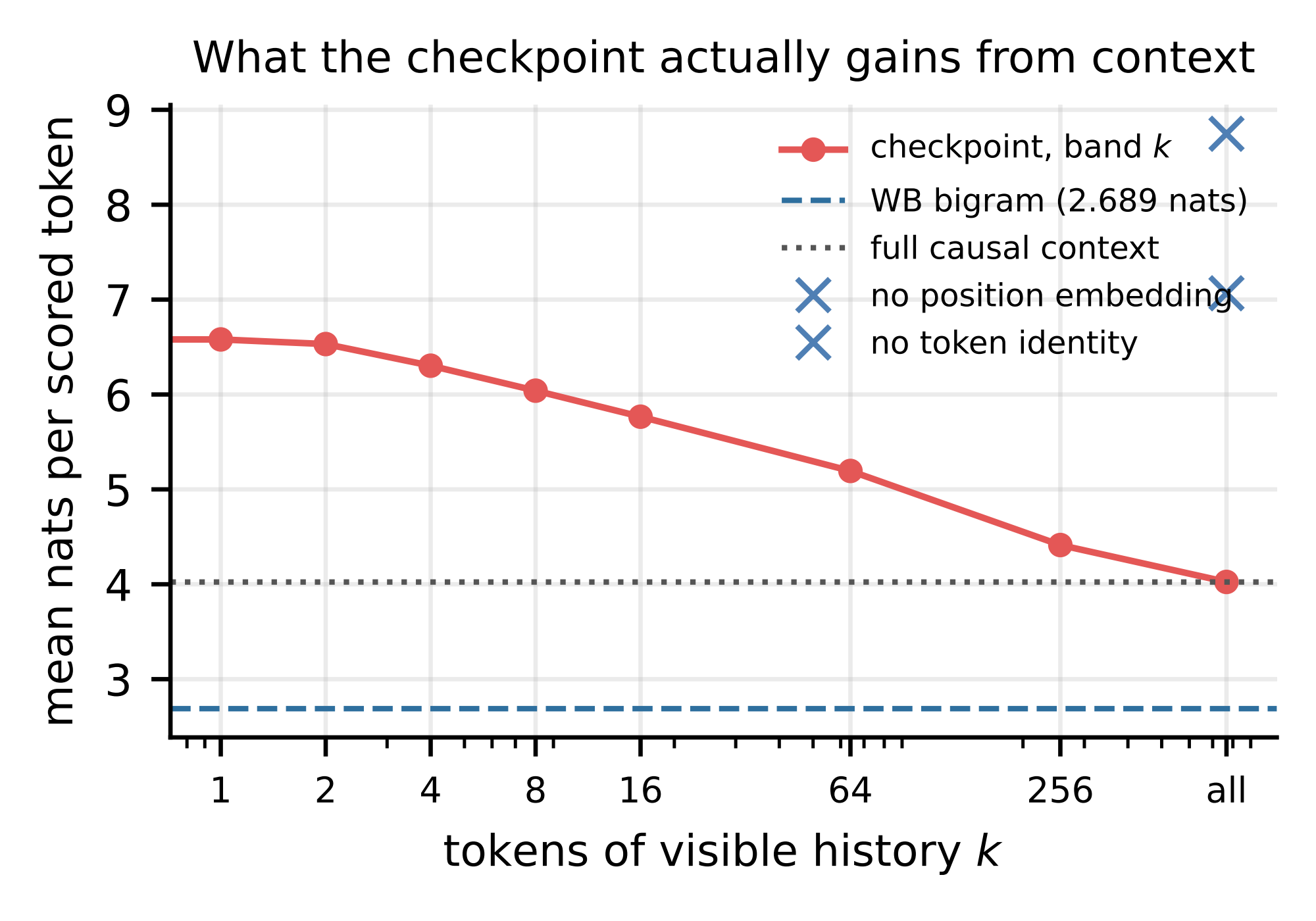

What the checkpoint actually gains from context
mean nats per scored token
9
8
7
6
5
4
3
checkpoint, band k
WB bigram (2.689 nats)
full causal context
no position embedding
no token identity
1
2
4
8
16
64
256
all
tokens of visible history k